Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 4, April 2017, pg.418 – 426

SURVEY OF DATA MINING TECHNIQUES FOR SOCIAL NETWORKING WEBSITES

D. Kavitha, MCA., M.Phil.,
Assistant Professor, KG College of Arts and science, Coimbatore

ABSTRACT: Social network has gained remarkable attention in the recent decade. Social network sites such as Twitter, Facebook, accessing them through the internet and the web 2.0 technologies has become more comfortable. People are more interested in and relying on social network for news, information and opinion of other users on diverse subject matters. This cause to generate massive data characterised by three issues namely; size, noise and dynamism. These often make social network data complex to analyse manually and results in the pertinent use of computational means to analyse them. Various data mining are used for detecting useful knowledge from massive datasets like trends, patterns and rules. These techniques were used in information retrieval, statistical modelling and machine learning. These techniques use steps like data pre-processing, data analysis, and data interpretation processes in the course of data analysis. This paper discusses different data mining techniques used in mining different aspects of the social network in the recent decades going from the historical techniques to the recent models, including our novel technique named TRCM.

Keywords: Social Network, Social Network Analysis, Data Mining Techniques

1. INTRODUCTION

Social network is a dedicated website enabling the user to communicate with each other by their post, videos, comments etc. Also they are web-based services that allow individuals creating public profile in a domain such that they can communicate with other users within that network. Social network has improved on the concept of technology of Web 2.0, by enabling the formation and exchange of User-Generated Content. Social network is a graph which includes of *nodes* and *links* representing the social relations on social network websites [1]. A *node* includes many entities and the relationships between them forming *links*. Social networks are important sources of online interactions and contents

sharing, subjectivity [2], approaches, evaluation , influences, observations , feelings, opinions and sentiments expressions bear out in text, reviews, blogs, discussions, news, remarks, reactions, or some other documents. Before the advent of social network, the homepages was popularly used in the late 1990s which made it likely for average internet users to share information. However, the activities on social network in recent times seem to have transformed the *World Wide Web (www)* into its intended innovative creation. Social network platforms enable rapid information exchange between users regardless of the location. Many organisations, individuals and even government of countries now pursue the activities on social network. The network enables big organisations, celebrities, government official and government bodies to obtain knowledge on how their audience reacts to postings that concerns them out of the enormous data generated on social network. The network permits the effective collection of large-scale data which gives climb to major computational challenges. Yet, the application of efficient data mining techniques has made it possible for users to discover valuable, accurate and three dominant disputes with social useful knowledge from social network data. Data mining techniques are capable of handling network data viz., **size, noise and dynamism**. The enormous nature of social network datasets require information processing to be done automatically for analysing within a reasonable time. Amusingly data mining techniques require huge data sets to mine the remarkable patterns from data, social networking sites appear to be perfect to mine with data mining tools. This form an enabling factor for advanced results for search in searching engines and also helps in better understanding of social data for research and organizational functions [3].

Data mining tools used for survey in this paper ranges from unsupervised, semisupervised to supervised learning. The rest of the survey are as follows. Section 2 examines the background of social networks. Section 3 lists research issues on social network analysis. Section 4 discusses some of the Graphical Theoretic tools used for social network analysis. Section 5 gives an overview of tools used for analyzing opinions conveyed on social network while Section 6 presents some of the sentiment analysis techniques used on social network. Section 7 describes some of the unsupervised classification techniques used in social network analysis**.** Section 8 present topics on detection and tracking tools used for social network analysis. The survey is concluded in Section 9 by stating the directions for future work.

2. SOCIAL NETWORK BACKGROUND

Social networking is the practice to expand the number of one's business and/or social contacts of one's by making connections through individuals, often through social media sites such as Facebook, Twitter, LinkedIn and Google+.

Social networks are meant for affordable and universally-acclaimed communication activities posting, product reviews online means. Networking websites pervades many aspects of our daily lives such as social interactions, news discovery, sharing, recommendations, emotions and feedback, professional networking, and social reinforcements. It is observed that more people are depending on the social network for information in real time.

Users at many times make decisions based upon the information posted by unknown individuals on social network making the increase in the degree of reliance on the credibility of these sites. Social network has succeeded in transforming the way different entities source and retrieve valuable information irrespective of their location. Social network has given the users the privilege to give opinions with very little or no restriction

2.1 SOCIAL NETWORK – POWER TO THE USERS

Social sites have undoubtedly bestowed inconceivable privilege for their users to access the readily available never-ending uncensored information. For example: Twitter, permits its users to post events in real time way ahead the broadcast of such events on traditional news media. Social network also allows the user to express their views, i.e., be it may be positive or negative[4]. Organizations are now mindful of the implication of consumers' opinions posted on social network sites to the credit of their products or services and the overall success of their organisations. These entities follow the activities on social network to keep side by side with how their audience reacts to issues that concerns about them [4]. Considering the enormous volume of data being generated on social network, it is important to find a computational means to filter, categorise, classify and analyse the social network contents.

3. RESEARCH ISSUES ON SOCIAL NETWORK ANALYSIS

A number of research issues that social networks face are as follows:

● LINKAGE-BASED AND STRUCTURAL ANALYSIS

link analysis is a data-analysis technique used to evaluate relationships between the nodes. Relationships may be identified among various type of nodes or objects, including organizations, people and transactions. Link analysis has also been used for investigation of criminal activity.

- Aggarwal, 2011.

● DYNAMIC ANALYSIS AND STATIC ANALYSIS

Static analysis are assumed to be easier to carry out than streaming networks. Here the social network changes gradually over time and analysis on the entire network will be done in batch mode. Conversely the dynamic analysis of streaming networks like Facebook and YouTube were very difficult to carry out than the former. Data generation using these networks are at high speed and capacity. The following sections present various data mining approaches used in analysing social network data.

4. GRAPH BASED K-MEANS CLUSTERING

Graph theory is probably the main method in social network analysis in the early history of the social network concept. K-Means algorithm is the simplest and most commonly used vector quantisation method. K-means clustering partitions data into clusters and minimises distance between cluster centers and data related to clusters. The approach is applied to social network analysis in order to determine important features of the network such as the *nodes* and *links*

4.1 COMMUNITY DETECTION USING HIERARCHICAL CLUSTERING

Community is a smaller compressed group [7] within a larger network. Community formation is said to be one of the important characteristics of social networking websites. Community detection method are able to assign a node not only to one community but also to many communities. Weights of all edges in complex networks are assumed to be the same in community detection methods. Communities on social networks, like any other communities in the real world, are very complex in nature and difficult to detect. Different authors have applied many clustering techniques to detect communities on social network with ***hierarchical clustering*** being mostly used. Most hierarchical clustering methods requires advance input. This technique is a combination of many techniques used to group nodes in the network to reveal strength of individual groups which is then used to distribute the network into communities. Hierarchical clustering inclues ***Vertex clustering method***, where graph vertices can be resolved by adding it in a vector space so that pair-wise length between vertices can be measured. Two peoples on social network having several mutual friends are more likely to be closer than two people with fewer mutual friends in the network.

4.2 SEMANTIC WEB OF SOCIAL NETWORK

The ***Semantic Web environment*** makes knowledge sharing and reusability possible over different applications and community edges. Discovering the evolvement of ***Semantic Web*** (***SW***) enhances the knowledge of the importance of ***Semantic Web Community*** and emphasizes the synthesis of the Semantic Web. The work employs the concept of ***Friend of a Friend (FOAF)*** to explore how local and global community level groups expand and change in large-scale social networks on the *Semantic Web.* The study revealed the evolution outlines of social structures and forecasts future lift. Likewise application model of ***Semantic Web-based Social Network Analysis Model*** creates the ontological field library of social network analysis combined with the conventional outline of the semantic web to attain intelligent retrieval of the Web services.

5 ASPECT-BASED/FEATURE-BASED OPINION MINING

Aspect-based also known as ***feature-based analysis*** is the process of mining the area of entity customers has reviewed. This is because not all aspects/features of an entity are often reviewed by customers. It is then necessary to summarise the aspects reviewed to determine the split of the overall review whether they are positive or negative. The sentiments expressed on some entities are easier to analyse than others, one of the reason being that some reviews are ambiguous. The aspect-based opinion problem lies more in blogs and forum discussions than in product or service reviews. The aspect/entity (which may be a computer device) review is either *'thumb up'* or *'thumb down',* thumb up life form positive review while thumb down means review negative. Conversely, in blogs and forum discussions both aspects and entity are not familiar and there are high levels of insignificant data which constitute noise. It is therefore necessary to identify opinion sentences in each review to determine if indeed each opinion sentence is positive or negative. Opinion sentences can be used to summarize aspect-based opinion which enhances the overall mining of product or service review.

6 HOMOPHILY CLUSTERING IN OPINION FORMATION

One way to find communities is to use the principle of homophily, which mean that two people tend to communicate more often if they share similar views. Using this phenomenon, those bloggers who have more edges within themselves can be considered as a community. Identifying a set of bloggers that communicate more often among them implies that they share similar views, opinions, or interest; hence they form a community. However, this approach to community discovery is purely based o network information. Opinion extraction identifies subjective sentences with sentimental classification of whichever positive or negative.

7 SENTIMENT ANALYSIS OF SOCIAL NETWORK

Sentiment analysis also called as opinion mining. The main aim of it is to define the automatic tools able to extract one-sided information from texts in natural language, such as opinions and sentiments, so as to create structured and actionable knowledge to be used by either a decision support system or a decision maker. Sentiment analysis can be referred to as discovery and recognition of positive or negative expression of opinion by people on diverse subject matters of interest. Depending on the field of application, several names are used for sentiment analysis(eg, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis and review mining). It is commendable of note that the enormous opinions of several billions of social network users are devastating, ranging from very important ones to mere assertions Consequentially it has become necessary to examine sentiment expressed on social network with data mining techniques in order to generate a meaningful frameworks that can be used as decision support tools. Diverse algorithms are in use to ascertain sentiment that matters to a topic, text, document or personality under review. The purpose of sentiment analysis on social network is to recognize potential glide in the society as it concerns the attitudes, observations, and the expectations of stakeholder or the populace. This recognition enables the entities concern to take prompt actions by making necessary decisions. It is important to decode sentiment expressed to useful knowledge by way of mining and analysis.

8 CLASSIFICATION OF SOCIAL NETWORK DATA

Artificial neural network is the mathematical model based on biological neural network. It consists of set of processing units which communicate together by sending signals to each other over a large number of weighted connections. Such, processing units are also called as neurons and they are responsible for receiving input from neighbours or cells or variables or external sources and using this input to compute an output signal which is propagated to other units. However each unit also adjusts the weight of connections. It is very necessary to evolve neural network by modifying the weights of connections so that they become more accurate. The neural network should be trained by feeding it teaching patterns and letting it change its weights. This is learning process. There are three types of learning methods:

- Supervised learning where the network is trained by providing it with input and matching output patterns.
- Unsupervised learning where the output is trained to react to clusters of pattern within the input. There is no a priori set of categories into which the patterns are to be classified.
- Semisupervised learning where the test nodes need to be predicted are known. --- - Reinforcement learning is the intermediate form between supervised and unsupervised learning.

8.1 SEMI-SUPERVISED CLASSIFICATION

Since the social network data usually come in huge sizes, in addition there are usually, a huge number of unlabelled instances. In such cases, it could be possible to use the information other than labels that exists in the unlabelled data, which leads to use of semi supervised learning algorithms. When the test nodes whose class will need to be predicted are known. In semi supervised learning, the unbalanced instances can be used to monitor the variance of the produces classifiers, to maximize the margin and hence to minimize the complexity.

8.2 SUPERVISED CLASSIFICATION

While clustering techniques are used where basis of data [8] is established but data pattern is unknown, classification techniques are supervised learning techniques used where the data organisation is already identified. It is worth of mentioning that understanding the problem to be solved and opting for the right data mining tool is very essential when using data mining techniques to solve social network issues. Pre-processing and considering privacy rights of individual should also be taken into account. Nonetheless, since social media is a dynamic platform, collision of time can only be rational in the subject of topic recognition, but not substantial in the case of network enlargement, group behaviour/ influence or marketing. This is because this attributes are bound to change from time to time.

8.3 UNSUPERVISED CLASSIFICATION

In Data mining, unsupervised learning tries to the find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled then there is no error or reward signal to evaluate a potential solution.

9 TOPIC DETECTION AND TRACKING ON SOCIAL NETWORK

Topic Detection and Tracking (TDT) on social network employs different techniques for discovering the emergent of new topics [9] (or events) and for tracking their subsequent evolvements over a period of time. *TDT* is delivery high level of attention recently. Many researchers and authors are conduct research on *TDT* on social network sites, especially on Twitter [5]; [6]. The main aim of TDT was to develop core technologies for news understanding systems. More specifically its tasks focused on discovering and keeping track of real world events in multi lingual news streams from various sources. Various methods have been residential for this task, including machine learning and query expansion based methods.

10 CONCLUSION AND FUTURE WORK

Different data mining techniques are used in social network analysis as covered in this work. The techniques are ranging from unsupervised to semi-supervised and supervised learning methods. As far, different levels of improvements have been achieved either with combined or solitary techniques. The result of the experiments conducted on social network analysis is supposed to have shed more light on the structure and activities of social networks. The varied experimental results have also established the bearing of data mining techniques in retrieve valuable information and contents from huge data generated on social network. Future survey will tend to investigate novel up to date data mining techniques for social network analysis. The survey will compare parallel data mining tools and suggest the most fitting tool(s) for the dataset to be analysed. The table also contain the approaches engaged, the experimental results and the dates and authors of the approaches.

REFERENCES

1. Borgatti, S P.: "2-Mode concepts in social network analysis." *Encyclopedia of Complexity and System Science*, 8279-8291, 2009.
2. Asur, S., and Huberman, B.: "Predicting the future with social network." *Web Intelligence Agent Technology (WIIAT), 2010 IEEE/WIC/ACM International Conference on*. Vol. 1. IEEE, 2010.

3. Aggarwal, C.: *An introduction to social network data analytics.* Springer US, 2011.

4. Castellanos, M., Dayal, M., Hsu, M., Ghosh, R., Dekhil, M.: U LCI: A Social Channel Analysis Platform for Live Customer Intelligence. In: Proceedings of the 2011 international Conference on Management of Data. 2011

5. Adedoyin-Olowe, M., Gaber, M., Stahl, F.: A Methodology for Temporal Analysis of Evolving Concepts in Twitter. Proceedings of the 2013 ICAISC, International Conference on Artificial Intelligence and Soft Computing. 2013.

6. Becker, H., Naaman, M., Gravano, L.: Beyond Trending Topics: Real-World Event Identification on Twitter. *ICWSM*, *11*, 438-441, 2011.

7. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(6), 734-749, 2005.

8. Batista, G., & Monard, M.C., (2003), An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence, vol. 17, pp.519-533.

9. Blei, D.M., and Lafferty, J. D. ―Dynamic Topic Models‖, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006